SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

19IT601– Data Science and Analytics

III Year / VI Semester

Unit 2 – DESCRIPTIVE ANALYTICS USING STATISTICS

Topic 8: Dimensionality Reduction, PCA

Dimensionality reduction

- In Machine Learning, the number of attributes, features or input variables of a dataset is referred to as its dimensionality.
- The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant.
- More input features often make a predictive modeling task more challenging to model, more generally referred to as the curse of dimensionality.
- Dimensionality reduction techniques exist to find a way to reduce higher dimensional information into lower dimensional information
- Dimensionality reduction refers to techniques that reduce the number of features or input variables in a dataset.

Example : classify whether the e-mail is spam or not

Dimensionality reduction

Importance of Dimensionality Reduction

- The performance of machine learning algorithms can degrade with too many input variables.
- Having a large number of dimensions in the feature space can mean that the volume of that space is very large.
- This can dramatically impact the performance of machine learning algorithms fit on data with many input features, generally referred to as the “curse of dimensionality.”
- A lower number of dimensions in data means less training time and less computational resources and increases the overall performance of machine learning algorithms.
- Dimensionality reduction avoids the problem of overfitting.
- Dimensionality reduction is extremely useful for data visualization
- Therefore, it is often desirable to reduce the number of input features.

Dimensionality reduction

Components of Dimensionality Reduction
There are two components of dimensionality reduction:

Feature selection: Feature selection is based on omitting those features from the available measurements which do not contribute to class separability. In other words, redundant and irrelevant features are ignored:

- Filter - use scoring methods, like correlation between the feature and the target variable, to select a subset of input features that are most predictive.

- Wrapper - fitting and evaluating the model with different subsets of input features and selecting the subset the results in the best model performance

Feature extraction: considers the whole information content and maps the useful information content into a lower dimensional feature space.

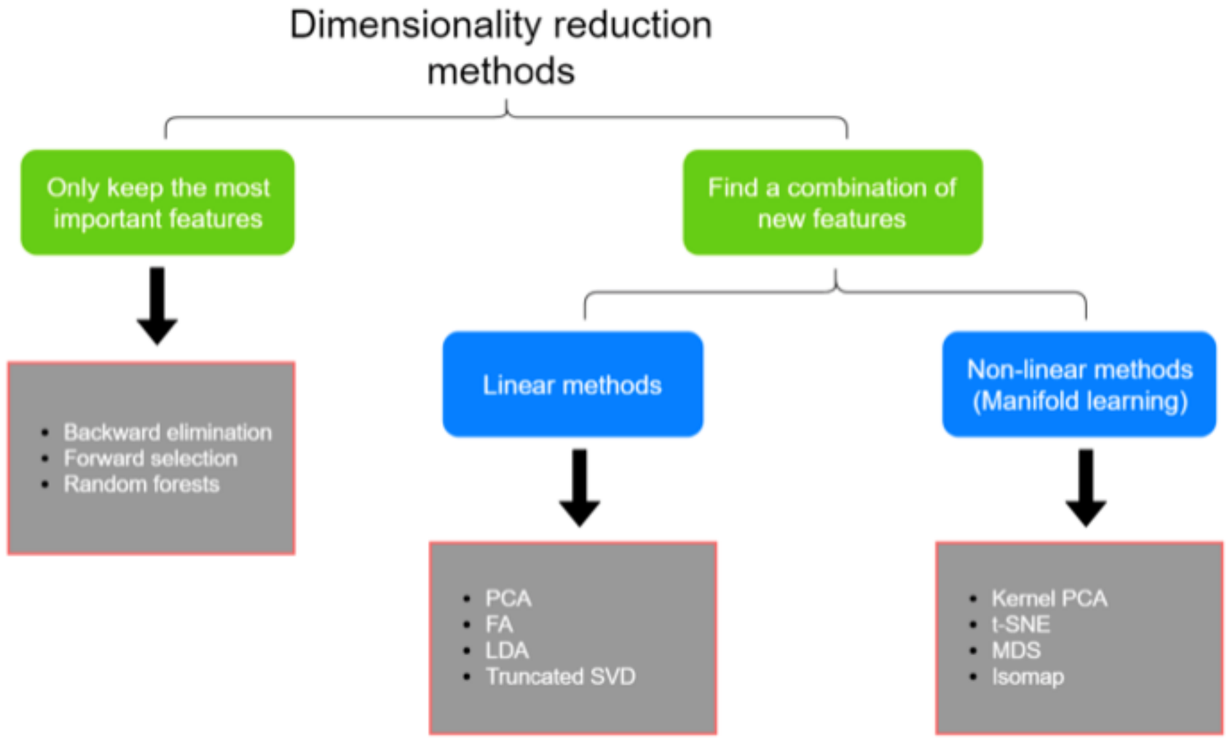
This reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions.

Dimensionality reduction

Methods of Dimensionality Reduction

The various methods used for dimensionality reduction include:

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)



Dimensionality reduction

Principal Component Analysis (PCA)

- This method was introduced by Karl Pearson.

- PCA is a linear dimensionality reduction technique (algorithm) that transforms a set of correlated variables (p) into a smaller k (k<p) number of uncorrelated variables called principal components while retaining as much of the variation in the original dataset as possible.

- In the context of Machine Learning (ML), PCA is an unsupervised machine learning algorithm that is used for dimensionality reduction.

- It works on a condition that while the data in a higher dimensional space is mapped to data in a lower dimension space, the variance of the data in the lower dimensional space should be maximum.

Dimensionality reduction

Principal Component Analysis (PCA)

PCA implementation is quite straightforward. We can define the whole process into just four steps:

1. **Standardization**: The data has to be transformed to a common scale by taking the difference between the original dataset with the mean of the whole dataset. This will make the distribution 0 centered.

2. **Finding covariance**: Covariance will help us to understand the relationship between the mean and original data.

3. **Determining the principal components**: Principal components can be determined by calculating the **eigenvectors and eigenvalues**.

4. **Final output**: It is the dot product of the standardized matrix and the eigenvector.

Dimensionality reduction

Goals of PCA

1.Extract the most important information from the data table.
2.Compress the size of the data set by keeping only this important information
3.Simplify the description of data set
4.Analyze the structure of the observations and variables.

In order to achieve these goals, PCA computes new variables called principal components, which are obtained as linear combinations of the original variables.

Dimensionality reduction

Linear Discriminant Analysis

Two criteria are used by LDA to create a new axis:

- Maximize the distance between means of the two classes.
- Minimize the variation within each class.

THANK YOU